CORE
Metadata, citation and similar papers at core.ac.uk
Provided by IntechOpen

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

7

Description and Publication of Geospatial Information

Arturo Beltran, Laura Díaz, Carlos Granell, Joaquín Huerta and Carlos Abargues
Universitat Jaume I de Castellón
Spain

1. Introduction

Information systems have evolved to service-oriented architectures (SOA) where dedicated desktop applications have turned into on-line data and services. On one hand this distributed environment let users to share (resources) data and tools, but on the other hand there is a need to develop mechanisms to allow users to find and access to these distributed resources efficiently.
Current trends for discover and access geospatial information are being addressed by deployment of interconnected Spatial Data Infrastructure (SDI) nodes at different scales to build a global spatial information infrastructure (Masser et al., 2008; Rajabifard et al., 2002) being the SOA paradigm in the geospatial domain.
However, current Geographic Information Systems (GIS) and the services provided by the SDIs fail to allow transparent navigation between related geographic information resources. In SDI like in other domains, metadata are a necessary mechanism to describe the information, and together with Catalogue Services are the key elements for discovery and information fusion possibilities (Nogueras et al., 2005; Díaz et al., 2007).
In this context, pointing out this need, there are directives such as INSPIRE[1], that at European level, mandates the creation and maintenance of metadata and related discovery services (Craglia et al., 2007), these elements are, often, the first visible elements of added value in SDIs.
Metadata allow us to describe data and, based on it, we could organize, publicize and facilitate the access to such information. Traditionally, it has been the user or the data provider who creates these metadata that will be published in Catalogue Services, for being discovered and accessed later by different users in a SDI. The fact of generating metadata like who created the data, where are they placed, etc. is a laborious task, fundamentally because the traditional metadata formats are large and complex, the users who are documenting the data usually have no knowledge about some metadata of the original data due to the lack of information supplied by the provider, etc.

[1] http://inspire.jrc.ec.europa.eu

In this sense, the production of metadata becomes a laborious job that consumes a large amount of time and effort becoming a task released into the background despite its major importance. This provokes, in reality, a scarcity in metadata availability in SDI and consequently difficulty in data discovery and a miss functioning SDI.
For all the above reasons there is a need to facilitate metadata production to easily create, with minimal user intervention, metadata descriptions when the data are created. In this way, data and metadata can be packed, forming a logical unit, created at the same time and minimizing the inconsistence between data and their metadata.
In this chapter we present a methodology for documenting geospatial information. This methodology provides mechanisms to automate the generation and publication of metadata. For demonstration purposes we describe a prototype implemented within an open-source software GIS/SDI client. This prototype is capable of semi-automatic extraction of explicit metadata from data resources, metadata edition and publication to be catalogued for data discovery in an SDI.
The nature of this integrated workflow that facilitates metadata creation and management, will hopefully contribute to a change in mindset as to the cost/benefit ratio of generating and exploiting metadata, a necessary ingredient for successful SDI.

2. Background

2.1 Geospatial Information

There are studies showing that most of the information (more than 80%) is likely to be linked to a geographic position. When we talk about geospatial information we are talking about data intrinsically related to a geographic position. Although there exists formats specially supporting geospatial data, any other data or information, not considered spatial in nature can be georeferenced and considered as such.
Georeferenced resources are then resources of any nature that have defined their existence in physical space. That is, those that have established their location in terms of map projections or coordinate systems. Nowadays, the act of georeference has gone beyond the fields of geoscience and GIS, thanks to the emergence of new tools which their ease of use has expanded and democratized this task outside of the current technical context.
The use of tools like Google Earth[2], Flickr[3], etc. has meant a qualitative leap in terms of georeferencing, extending the use of georeferencing resources traditionally limited to geodata in geosciences and GIS specialists, and thus accelerating the emergence of a geosemantic web, (Cerda, 2005). In the same way, the overcrowding and constant evolution of the georeferentiation has been boosted by the use of mashups in Web 2.0 sites, allowing the location of digital content (photo, video, news, 3D models, etc.) in digital mapping, nowadays called neogeography (Goodchild, 2007) (Goodchild, 2008).
All this georeferenced content, like geospatial data, can be described by using metadata and published in Catalogue Services in order to be integrated in SDI.

[2] http://earth.google.es
[3] http://www.flickr.com

2.2 Description of Resources

A description is the explanation, in a detailed and ordered way, of how is certain person, place, object or anything, through the explanation of its various parts, characteristics or circumstances.

As we said earlier, metadata allows us to describe data and, based on it, we could organize, publicize and facilitate the access to such information. Metadata are commonly defined as "structured data about data" or "data that describe the attributes of a resource" or simply "information about data". In other words, metadata is the information that describes the content, quality, condition, origin, and other characteristics of data. Metadata is the information and the documentation that enable data to be well understood, shared and used effectively by all types of users over time.

These metadata or data description must be generated according to a standard in order to fulfil the minimum requirements for interoperability. One of these metadata standards is DublinCore (DC), this standard was born originally to describe Web resources in a general way proposed by the initiative "Dublin Core Metadata Initiative" (DCMI)[4]. This initiative, created in 1995, promotes the dissemination of interoperable metadata standards and metadata vocabularies to build more intelligent information search systems. The DC standard has been approved as an American standard (ANSI/NISO Z39.85), in the technical European committee CEN/ISSS (European Committee for Standardization / Information Society Standardization System) and since 2003 also as an international standard by ISO (ISO 15836:2003 "Information and Documentation - The Dublin Core Metadata Element Set").

The need of this kind of metadata standards is pointed out by organizations like World Wide Web Consortium (W3C)[5]. There are many other standards utilized for specific domains, for example, we can find various metadata formats for multimedia resources, like: Apple ITUNES XML, Yahoo MediaRSS, Cablelabs VOD Metadata Content Specification 2.0, MPEG-7 standard, W3C SMIL Standard, etc. W3C tries to standarize all these metadata formats and provide a way to work efficiently. (Toebes, 2007).

The W3C proposal is to develop metadata extending languages based in XML (eXtensible Markup Language) (Bray et al., 2000) or RDF (Resource Description Framework) (Manola y Miller, 2004). In this way in the geospatial domain there is a general consensus.

As we mentioned in the section before, potentially, any resource could be georeferenced and be integrated with other geospatial information in or outside SDI environments. As we focus on a methodology for description of geospatial information, we describe next the goals of the geographic metadata creation and the standards used in this domain.

Geographic metadata help people involved in the use of geographic information to find the information that they need and determine how best to use them (Nebert, 2004). In (FGDC, 2000) it is stated that the creation of geographic metadata has three major goals (which are also benefits):

- Organize and maintain investments in data made by an organization: Metadata seek to promote the reusability of data without having to turn to the team that was responsible for its initial creation.

[4] http://www.dublincore.org

[5] http://www.w3.org/

- Publicize the existence of geographic information through catalog systems: Metadata records are usually published through catalog systems, sometimes also referred as directories. Electronic catalogs not differ too much from the traditional library catalogs except for the fact that it offers a standardized interface for search services. Thus, these catalogs are the tool that put consumers in touch with the producers of information. By means of the publication of geographic information resources through a catalog, organizations can find data to use, other organizations with who share data and maintenance efforts and customers for these data.
- Facilitate the access to the data, their acquisition and a better utilisation of the data achieving information interoperability when it comes from various sources: Metadata help receiving users or organizations in the processing, interpretation and storage of data in internal repositories.

Within the world of geographic information have been defined recommendations for the creation of metadata, whose main purpose is to provide a "hierarchical and concrete" structure to describe fully each of the data to which they refer. These recommendations have been created and approved by standardization bodies according to opinions of experts in the area. These recommendations, in form of standards or metadata schemas, provide criteria to characterize their geographic data properly.
Throughout the years have emerged, at national or European level, even within a specific domain, a set of initiatives to standardize the creation of metadata. However, these initiatives have been repealed for harmonization with the international standard ISO19115:2003[6]. Even the new version of the American standard CSDGM[7] will converge with the international standard.
Regardless of the metadata standard used, it is usual to classify the elements of metadata respect on their role within the paradigm "discovery, evaluation and access" established in (Nebert, 2004):

- Discovery metadata elements are those that allow minimally describe the nature and content of a resource. These elements usually respond to the questions "What, Why, When, Who, Where and How". Typical elements in this category would be the title, the description of the data set or its geographic extension.
- Exploration metadata provide information that allow verify that the data are in accordance with the desired purpose, assess their properties or contact with the organization that will provide further information.
- Exploitation metadata include those necessary descriptions for access, transfer, load, interpret and use the data in the final application in order to be exploited.

Another important aspect related to the metadata schemas is their level of detail, which is defined by the choice of the standard itself and the creation of special extensions and profiles. First, the chosen standard defines a more or less large set of elements with different condition: mandatory, optional and mandatory if applicable or conditional. An extension of the standard usually consists on adding new constraints (e.g. conversion of optional elements to mandatory), extension of code lists and the creation of new elements and

[6] http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020
[7] http://www.fgdc.gov/metadata/csdgm

entities. Some standards such as ISO19115:2003 and CSDGM provide methods for the extension of the metadata within their specification. And if there are a big number of these extra features (they involve the creation of a considerable number of elements), ISO19115:2003 recommends making a formal request for the creation of a specific application profile for that community of users who require it.
However, although the specific profiles and the optional and conditional elements facilitate certain flexibility to the geographic metadata, most of the common used standards like CSDGM and ISO19115:2003 are too complex (Nebert, 2004), both define more than 350 elements distributed into multiple hierarchical sections. This complexity means that, to complete the geographic metadata, it is necessary to devote a big amount of time and highly qualified human resources.
Automatic mechanisms for generating metadata in standard format would be a helpful way to assist user to increase the number of available metadata in distributed environments improving the discovery of the data in an efficient way.

2.3 Generation of Metadata

Metadata is usually created by data providers, generated manually and stored (separated from the resource) in catalogs, according to digital libraries tradition, to be found later for informational purposes. However, practical problems with their creation and maintenance are limiting their effectiveness for tasks such as discovery and evaluation of the usefulness of a given resource.
Some authors emphasize as causes of this low effectiveness the complexity of the rules and standards in the geospatial context or the low automation and synchronization between the creation of data and metadata. In terms of complexity, (Bodoff et al., 2005) regret the overhead of planned uses for some metadata: according to certain rules some metadata must provide at the same time the documentation, the configuration and the access point to the resource. Other authors point out the necessity to automatize data generation, (Bulterman, 2004; Manso et al., 2004).
Nowadays metadata are usually created manually, and only few of them are extracted automatically by software, for example, geographical extent or the date of creation. Although theoretically only must be introduced by hand subjective descriptors such as the abstract, but the complexity and variety of formats limit the application of automated techniques.
Due to the increasing need of metadata to find the great amount of data available in distributed environment, especially in geospatial information systems, being deployed as SDI, there are numerous software applications that try to facilitate this metadata generation. Most of these application started supporting CSDGM standard and ISO 19115. There is a good survey on these applications available in the FGDC metadata working group[8].
The purpose of this section is to make a small state of the art of existing proposals to improve the automated generation of metadata. The hypothesis is that the automatic generation of meta-information permits decrease human interaction in the creation of metadata, reducing the associated workload and the obstacles arising from the complexity of the metadata schemas that metadata creators must face.

[8] http://www.fgdc.gov/metadata/iso-metadata-editor-review

2.3.1 Methods aimed at the extraction

Actual models of representation of georreferenced information, especially the raster and vector spatial representation models, are characterized by being highly structured and are manifested in multiple exchange formats. Due to the complex nature of digital resources, it is not possible to effectively reuse methods for automatic generation of metadata already existing in the context of information retrieval on textual type documents (e.g. Web browsers). On the other hand, the few existing GIS tools that offer automatic deduction of metadata for raster and vector formats are based on the analysis of these specific formats and the implementation of ad-hoc mechanisms that process the data in these formats to extract information which is used later to populate the metadata elements (Manso et al., 2004).

Among the applications that perform an automatic extraction of metadata from certain geospatial data exchange formats is the free software tool CatMDEdit[9] (Zarazaga-Soria et al., 2003). As reflected in the work done by (Manso et al., 2004), the amount of information that can be extracted depend fundamentally on the representation model used, and its own file format. In this way, there are elements that can only be extracted from certain types of data and files, while others, such as the size of the data, could be obtained in any circumstances.

Another well-known tool and widely used that includes automatic metadata generation functionality from geographic data is ESRI[10] ArcCatalog, available from version 8 ESRI ArcGIS. This tool allows the automatic loading of a number of basic fields and the synchronized update of data and metadata. To improve this tool have been created some extensions, such as the Metadata Editor of the *Núcleo Español de Metadatos* (NEM), it is a fully integrated tool with the ArcCatalog application, capable of generate a metadata record that meets the standard ISO19115:2003 and NEM v1.0[11]. Metadata created with this editor will be integrated with the ArcCatalog metadata search functionality, as having been generated by the application (Sanchidrian & Calle, 2005).

Regarding to other data formats, such as text, sound or video documents, content creation software, that is, the range of programs used to create these resources, usually support some automatic metadata generation from the content that they generate. For example, MS Office attaches to the document a title based in the text of its first line, apart from other technical metadata such as dates of creation or modification and the author information. These metadata created by the content creation software are often used by the file system to index and sort the contents. All this kind of metadata can be collected during the creation process (Greenberg et al., 2005), but during the creation of digital content there are other metadata that can be automatic generated, they are usually used in various visualization applications, but not usually taken into account for the description of the resource for future discovery.

A couple of examples of the type of work that is being developed in this area can be the report of (Greenberg et al., 2005) on the generation of metadata for MS Word, Acrobat, Dreamweaver, CityDesk, WinAmp… file formats or the DCS (Dublin Core Services) project, which develops a set of services and applications for the automatic metadata extraction from more than 10 types of digital formats (XML, BibTex, XHTML, PNG, etc.), this project aims to support the development and widespread application of the Dublin Core standard format.

---

[9] http://catmdedit.sourceforge.net

[10] http://www.esri.com

[11] http://www.idee.es/resources/recomendacionesCSG/NEM.pdf

2.3.2 Methods aimed at the inference

We can infer metadata from other metadata or from geodata, using various techniques of data mining, data recovery, using the context surrounding the data, reasoning techniques and so on. We could know the administrative limits of a geo-spatial data from the knowledge of their bounding box using a gazetteer, or maybe we can infer a more or less adequate abstract from the information of the name, the legend of a layer, its geographical position, etc.

In this sense, (Taussi, 2007) proposes a metadata extraction based on three fundamental steps. The first step consist on apply some metadata extraction techniques largely based on the specific exchange format of the geographic data. Next step is the automatic deduction of the information regarding data quality, using brute force, stochastic or comparison techniques of the analyzed data with other reference data. Finally, the last step to apply is based on the utilization of data mining techniques that lead to obtaining a higher degree of knowledge about the data. Among the proposed data mining techniques, the following can be highlighted (Hand et al., 2001):

- Exploratory data analysis: Goal is to explore data without clear ideas of what we are looking for.
- Descriptive modelling: Idea is to describe all of the data. For example, showing the distribution of the data, partitioning of the data into groups or making models that show relationships between variables.
- Predictive modelling: Goal is to build a model that permits one variable to be predicted from the known values of other variables.
- Discovery methods: These methods are based on pattern detection, and idea here is to identify patterns, rules, outliers or combinations of items that occur frequently in data.
- Retrieval by content: It is based in the comparison of the contents of the dataset according to the pattern of interest to find similar patterns.

A work that tries to take a further step in the induction of metadata from the analysis of data is that developed by (Klien & Lutz, 2005). They propose a method for automatic annotation of geodata that consists of two main steps. In a first step, ontologies are defined from the definition of concepts (eg. floodplain) for a possible dataset. Depending on the spatial relationships that exist and should be verified with a reference dataset, it is checked whether the dataset corresponds to a floodplain, they check their connection to a nearby river, the altitude with respect to this, and if it is a flat terrain. In a second step, existing topological relationships are verified by a spatial processing for each type of relationship included in the concept, and if it meets all the dataset is semantically annotated with the concept. This approach requires a previous readiness to define concepts based on spatial relationships that makes the method is not directly applicable to any set of data. But in any case, it is helpful to specify formally the spatial analysis that allows checking whether a dataset meets certain characteristics, for annotate semantically.

Outside the geographical scope, there are also other works that exploit the idea of extracting information about other resources using techniques related to data mining. In this line we can mention the work done by (Kawtrakul & Yingsaeree, 2005) which provides a framework for extracting metadata from electronic documents, such as text documents or images; the study

of (Day et al., 2007) for extracting metadata of publications from the bibliographic references, or the articles of (Boutell & Luo, 2005) and (Suh & Bederson, 2007) where photographs analysis techniques (based on clusters) are presented to identify, for example, if the photograph was taken during the day or night, in a natural reserve or in a city, inside a place or outside...

2.4 Publication of Georreferenced Resources

The publication is the effect of revealing or expressing to the public some information, that is, the activity of making information available for public view.

In the world of geographic information is widely accepted that publishing means to make available to users some information of the data in a catalog service, as it is driven by Open Geospatial Consortium (OGC)[12] and its standards. However, OGC standards are not the only mechanism for publishing and searching geographic information or georreferenced resources. In the case of entities of medium or small size it might be appropriated to turn to simpler mechanisms that allow the content availability online. Z3950 standard (ISO23950[13]), widely used in digital library environments, includes a GEO5 profile that allows to extract Z3950 metadata as XML whose content is based on FGDC standard.

A more general mechanism, but whose philosophy and operation can be adapted to the field of the georreferenced resources is the Open Archives Initiative (OAI)[14]. This initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. OAI has its roots in the open access and institutional repository movements. Over time, however, the work of OAI has expanded to promote broad access to digital resources for eScholarship, eLearning, and eScience.

OAI provides us with the specification Open Archives Initiative Object Reuse and Exchange (OAI-ORE) that defines standards for the description and exchange of aggregations of Web resources. These aggregations, sometimes called compound digital objects, may combine distributed resources with multiple media types including text, images, data, and video. The goal of these standards is to expose the rich content in these aggregations to applications that support authoring, deposit, exchange, visualization, reuse, and preservation. Although a motivating use case for the work is the changing nature of scholarship and scholarly communication, and the need for cyberinfrastructure to support that scholarship, the intent of the effort is to develop standards that generalize across all web-based information including the increasing popular social networks of "Web 2.0".

The specification of standards for the publication of georreferenced information, whose implementation is feasible from a technical and economic point of view, results essential to the progress of technology and services.

[12] http://www.opengeospatial.org

[13] http://www.loc.gov/z3950/agency

[14] http://www.openarchives.org

3 Proposed Methodology

3.1 Metadata generation

The proposed methodology, to automatically generate complete metadata and of reasonable quality, avoiding as much as possible the participation of the user, is a combination of the methods described before orchestrated efficiently.

Initially, we start by obtaining all relevant information that can be extracted from the data resource itself, for example, the data size or the creation and modification data. Later, we try to extract as much information as possible from the data content. We must emphasize that this is one of the most important sources of information, so we must pay special attention on it. Moreover, how to analyze the resource and the amount of information available will depend entirely on the nature of the resource and the format in which they are represented. By this method we can extract explicit information in the data, for instance, in an email is easy to find information such as the sender, the recipient or the date that it was sent.

Now, we will add the common information pertaining to the creation and the exploitation context. From the creation context we can obtain relevant information as the organization or the company responsible of the data and the theme of the data. We can operate in a similar manner with the exploitation of context, obtaining information such as the theme or the resource quality offered by certain company. All these information can be previously set or revised by the user or automatically obtained exploring the data set and their context.

The next step is to consider collecting information from the process of creation of the data, obviously if it exists. It should be noted that this source of information is "volatile" due to the fact that it is only available at the time that data is created and for this reason we must collect and store all possible information at that moment. We consider that the information that can be obtained from the process of creation of the data is very important and rarely taken into account. By this method we can accurately find out relevant information such as the creation process in order to replicate the results later, costs (computational, temporal, economic, etc.) or the author of the data. In addition to the information that we had commented just now, during the process of creation of the data, a sensor or other measuring mechanism can provide relevant information. Some magnitudes could be measured such as elevation, position or temperature, and incorporate this values to the metadata automatically. We can find a good example of this method in some digital cameras that use it to add, among others, the information that provides their integrated global positioning system (GPS)[15] device to the images in the form of EXIF[16] tags.

Having reached this point we already have a base of information, and applying to it some deductive methods we will try to extend it. One way to deduce new metadata is that an element of metadata is created through a direct correspondence with another existing metadata element. For example, you can get the place name corresponding to the data using the 4 coordinates of their bounding box, using a gazetteer service. Another way to deduce new metadata is based on the calculation of a new element of metadata through a computation process of the data themselves. In this sense there are many lines of investigation open that cover a wide range of possibilities. We can find from different techniques to analyze/process text documents or web pages to find out its main theme or

[15] http://en.wikipedia.org/wiki/Global_Positioning_System

[16] http://www.exif.org

keywords, to other techniques that employ the geodata themselves, for example, to determine the province of a town by topological calculations. The last deductive method is the inference of metadata from other existing metadata or from the data content. It represents the best method, in fact in some situations the only one applicable, for the post-hoc creation of metadata, that is, document existing geodata. An example would be to infer the season of the data using the temperature metadata element, perhaps obtained by measurements such as we have explained above, so a rule would establish for a temperature below 15 degrees in Tenerife that we can suppose winter. Today it is obvious that the creation of this inferred metadata overlaps extensively the research fields of data mining and data recovery (Goodchild, 2007).

Finally, we must never forget to offer the user the possibility of modify or introduce the information, although the idea is that users increase their confidence in the methodology in base of the observation of its acceptable results and tend to not participate in the metadata generation process.

While most methods are applicable throughout the lifecycle of data, other methods are only applicable at the moment that data are created. We must pay special attention on them since most times the information that is not collected at that time is lost forever, and some of this information can be essential for a correct description of the resource.

The proposed methodology will allow the improvement of the automatic generation of metadata and their quality, collecting information that is currently ignored, such as that one that is coming from the creation process that can be very important to describe resources properly. Consequently, the result of applying this methodology will obtain more metadata, of higher quality and correction, with reduced participation of the user.

3.2 Metadata publication

Metadata publication is the second step of the proposed methodology. Once the metadata has been generated, users will be assisted in the automatic publication of this metadata. We can give users the possibility of publish these metadata automatically in an integrated way in the workflow. Hopefully, this will lead to increase the amount of published metadata since we are drastically reducing the necessary effort to correctly describe resources (by generating metadata manually) and to publish them.

In our methodology, metadata publication means to publish metadata in a Catalog Cervice, we will use the CSW[17] protocol, specifically the transactional profile (CSW-T) according to the OpenGIS Catalog Services Specification (Nerbert & Whiteside, 2004). This protocol supports the ability to publish and search collections of descriptive information (metadata) about geospatial data, services and related resources, so it covers our needs perfectly. Furthermore, it had become an OGC standard so it will be widely adopted by GIS applications.

However, we want to explore other ways to publish metadata and resources in order to obtain better levels of the capacity to be found. One way is publish the information resources directly in servers or social networks that support this kind of information using the metadata to document it properly in the server. For example, we can publish maps in MapServer[18], photographs in Flickr or GPS tracks in WikiLoc[19]. Other way to explore is put the resources

[17] http://www.opengeospatial.org/standards/cat

[18] http://mapserver.org

[19] http://www.wikiloc.com

available in order to be indexed by Internet search engine's bots, in this way we can try some techniques from simply put the resources available to build an associated KML with the metadata. Other way can be the use of peer-to-peer (P2P)[20] (Rüdiger, 2002; Antoniadis & Le Grand, 2007) networks to share data inside an organization network or globally. The main idea is to explore and test these alternatives and others in combination with various levels of metadata generation to measure the capacity to be found of the published resources.

4. Architecture

This section describes the architecture of our proposal to develop this methodology. The following architecture shows the modules and the connections to design an application that we have called *GeoCrawler*.

A crawler is an application that explores the content of a system in a methodical and automated way. This kind of applications is used to build an index the resources found in the system, in basis to the information extracted of each resource in its processing. In this way, *GeoCrawler* will explore the local machine (in the future we can consider to modify it to allow network exploration) and will try to generate metadata of the available geospatial information resources and later publish them according to their respective metadata.

To implement and fulfil the requirements of the proposed automatic metadata generation methodology we have decided to use a three-tier architecture (Eckerson & Wayne, 1995), to place some modules designed to implement the required functionality in each tier. This architecture is a client-server architecture in which the user interface, functional process logic ("business rules"), computer data storage and data access are developed and maintained as independent modules, often on separate platforms.

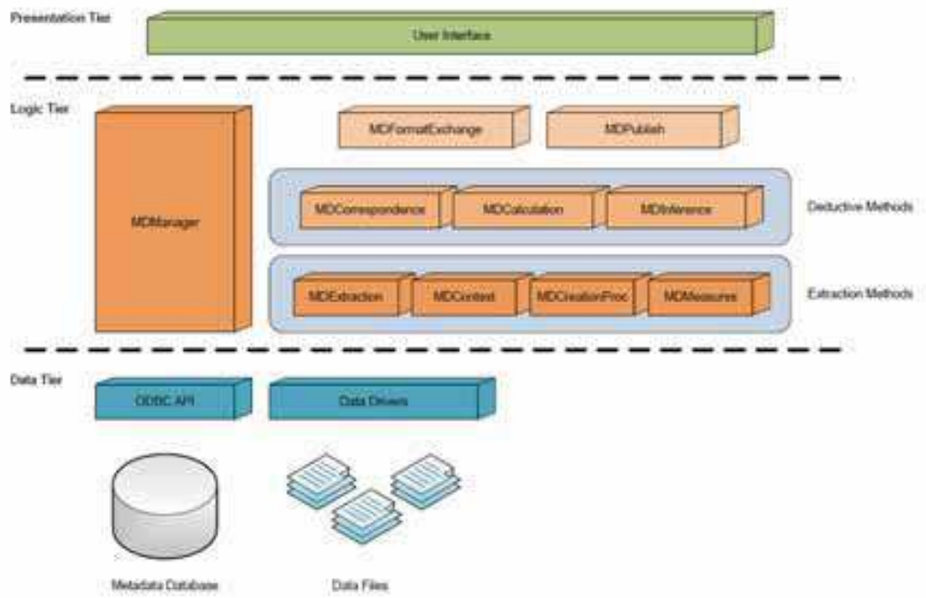

Fig. 1. General Architecture

[20] http://en.wikipedia.org/wiki/Peer-to-peer

As we can see in Figure 1, at the bottom of the figure and the lowest level of the application is the Data Tier, this tier consists on a database to store the generated metadata and the data files themselves. In this tier the information is stored and retrieved, so it must provide well defined interfaces to manage the data. In our case, this access is provided by data drivers to access to the data files and ODBC[21] to access to the database. This kind of design, keeps data neutral and independent from business logic, and also improves scalability and performance.

The next tier, which lies just above the Data Tier, is the Logic Tier. It controls the application's functionality by performing detailed processing. In the bottom level of this tier we can find the components implementing the metadata generation methods aimed at extraction. These components correspond to the methods described in the proposed methodology section. Thus, the *MDExtraction* module implements the extraction of all relevant information that can be extracted from the data resource and its content. The *MDContext* module will try to obtain all the information pertaining to the creation and the exploitation context. In a similar manner, the function of the *MDCreationProc* module is collect information from the process of creation of the data. Additionally, the *MDMeasures* module can acquire relevant information from sensors or other measuring mechanisms. At a higher level, based on their previous results, we can find the components implementing the deductive metadata generation methods. These methods will be the deduction of new metadata based on a direct correspondence with another existing metadata element (*MDCorrespondence*), the calculation of a new element of metadata through a computation process (*MDCalculation*) and the inference of metadata (*MDInference*) that includes data mining and data recovery techniques. We have to emphasize that new modules implementing new automatic metadata generation methods could be added.

According to the methodology, and as we can see on the architecture, on the top level of this tier and connecting to the modules which generate metadata, we have the *MDFormatExchange* responsible of generate standard formats and handles the transformation between them. At the same level of this module we have the *MDPublish* module that using, normally the metadata in any standard format, implements the publication business logic, publishing the data in a Catalogue Service or in any other way decided by the user. Finally, in this tier, and covering the whole layer scope, we can see the *Metadata Manager* component whose functionality is to orchestrate the metadata generation efficiently, provide the generated metadata to other components and offer the visible interface to the upper tier.

In the top of the Figure 1 we have the highest application level, where we find the Presentation Tier, this tier displays the information provided by the lower tiers through a graphical user interface. This user interface, moreover, allows users to interact, configure and operate with the application.

This kind of architecture, benefits from the advantages of modularized software containing well-defined module interfaces, it intends to allow any of the three tiers to be upgraded or replaced independently. This is very useful if we want to reuse some components (even the module containing the two lower tiers) and integrate it in other system, to incorporate the functionality of automatic metadata generation and management to any new or existing application.

[21] http://en.wikipedia.org/wiki/Open_Database_Connectivity

5. Case Study: Metadata Management Platform in gvSIG

We describe next a case of study in which we have implemented a proof of concept for our methodology and architecture. In this sense, we have implemented a prototype of the metadata manager, using the functionality and the extension possibilities that offers an open-source software GIS/SDI client called gvSIG[22].

We have extended gvSIG to facilitate, with an integrated workflow, metadata creation, management and publication. This prototype interacts with the gvSIG core to handle the metadata associated to all the resources pointed out to be described with metadata, and provide automatic extraction of explicit metadata from data resources for both internal metadata for user efficiency purposes and external metadata to be catalogued for data discovery in an SDI. In this case, with this integrated solution, we could get lots of information available in the process of data creation. The metadata manager will be working in the background annotating all the metadata while gvSIG users are working with their geospatial data, when it is required the metadata manager will use the implementation of the proposed metadata generation methodology to obtain as much information of the resource as possible, thus, without user interaction. As an added value gvSIG will be using these metadata, as internal metadata to avoid task duplication or recalculations and to visualize the resources properly. On the other hand, when the user wants to share data in a distributed environment like an SDI, he or she can use this metadata to publish metadata. The metadata manager will allow the user to visualize and edit the metadata according to one chosen standard format, and warn about the status of the metadata, for example to fulfil a minimum required set of elements of a certain standard format. Finally, included in this workflow, the prototype includes a user-friendly wizard to guide the user to publish these metadata in a Catalogue Service.

This prototype became available in October 2008 as a pilot plug-in of gvSIG. To sum-up it includes the metadata manager capable of semi-automatic extraction of explicit metadata from data resources for internal use or for being exported to a standard format and/or published in a catalogue service.

In this prototype we support GeoNetwork Opensource[23], as an implementation of the OGC CS-W because is one of the most popular and extended open source Catalog Service implementation. The prototype architecture is shown in the next Figure 2.

[22] http://www.gvsig.gva.es

[23] http://geonetwork-opensource.org

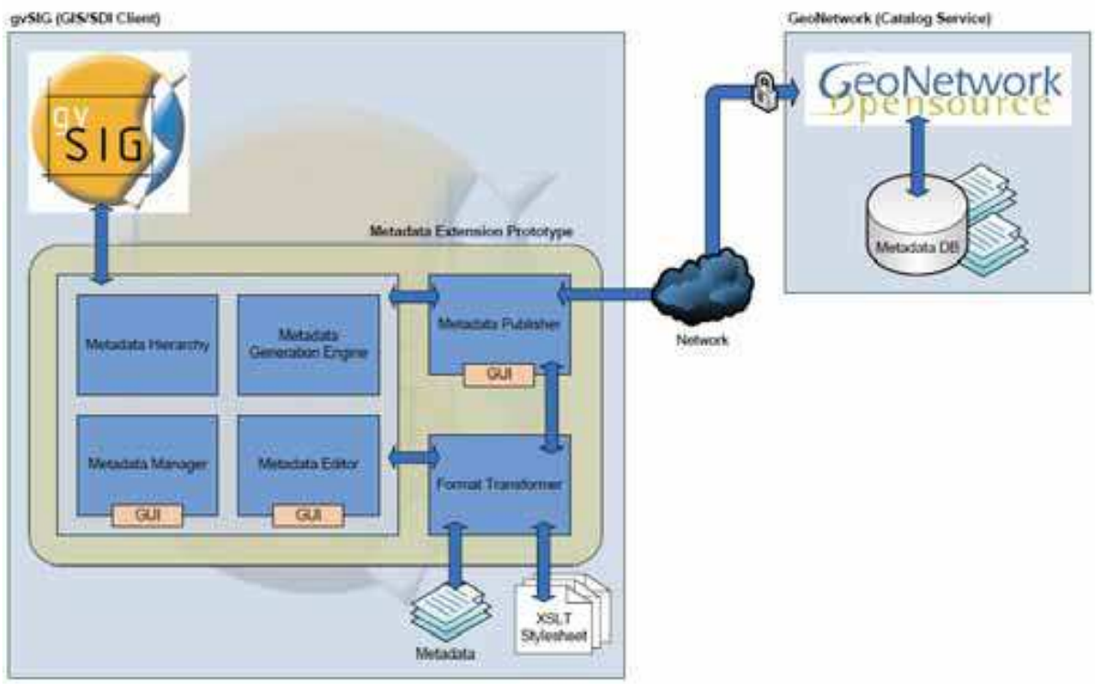

Fig. 2. Metadata Manager Prototype Architecture

As we can see in the Figure 2, in this prototype we define, within the central structure (or core) of gvSIG, an internal metadata dynamic object which would keep all types of metadata associated with their respective resource, as we can see reflected in the *Metadata Hierarchy* module .

The various metadata elements collected will be stored in an XML format file that will be saved together with the data for future uses. When a resource is created, thus it does not have associated metadata yet, the *Metadata Generation Engine* module will be used to generate all the possible metadata according to the proposed metadata generation methodology. In this prototype, we automatically extract so-called explicit metadata of the resource (format, resolution, spatial reference system, creation date, etc.) using the operating system information, and data drivers of gvSIG which are able to read file format headers and other information to collect metadata. As future work we will include inference and information retrieval techniques to create metadata according to the proposed metadata generation methodology, so the user will hardly have to edit or add metadata to publish it in an SDI, thus facilitating the proliferation of metadata and thus the resource discovery in distributed information platforms.

The Figure 3 shows a screen shoot where we can see part of the modules containing GUI (Graphical user interface) shown in the prototype architecture. These modules let the user to visualize and edit these associated metadata by using the metadata editor, he or she can add additional information (such as an adequate title or abstract) that might be required by the standard metadata formats. In this case, when a user wishes to edit the metadata to export it or publish it in a catalogue service, he will choose one of the supported standard formats for this purpose. Once it has been chosen, the metadata manager will start a wizard that will guide the user to view and edit the metadata according to the selected format and validating

the metadata fields. This wizard will guide users to import and export metadata too, validating it according to a standard format to be shared by multiple users without having to publish it in a catalogue service. As we see in the figure the *Metadata Editor* allows users to complete and verify the metadata record according to the selected metadata standard format.

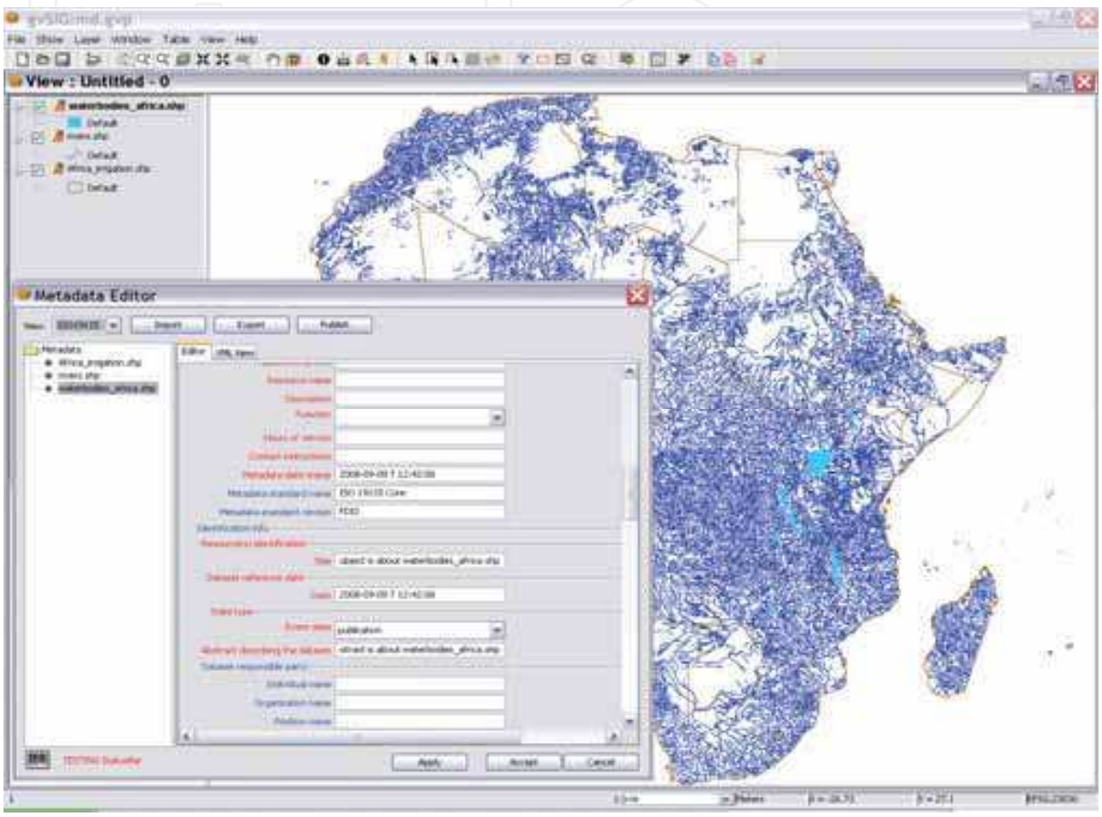

Fig. 3. Screenshot of the Metadata Extension Prototype (Metadata Editor)

As we see in Figure 3. This user interface also links with the *Metadata Publisher* module that will assist the user with a Publish wizard to publish this metadata in a Catalog Service to share the data in an SDI. In the next figure we can observe a screenshot of this Publication Wizard after having finished successfully.

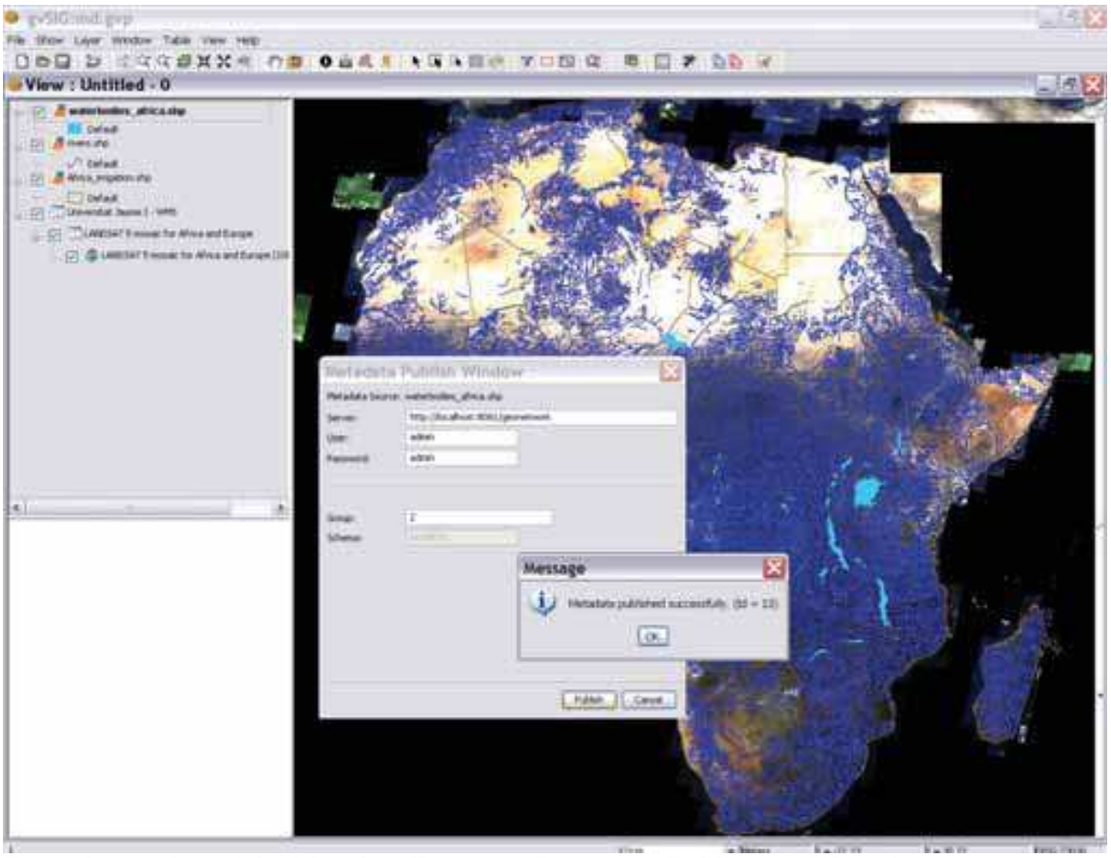

Fig. 4. Screenshot of the Metadata Extension Prototype (Publication Wizard)

Given the above information, let us look at a typical use case. A technician using gvSIG has combined basic geospatial data including terrain data such as slope and aspect, with vegetation data, to create a rough forest fire risk map. Assuming he or she has permission to share this new dataset, he then undergoes the process of publishing the risk map to a map server, and would also like to (or should be required to) publish its description to a metadata catalogue service such as that currently available at the European Commission INSPIRE Geoportal[24].

In our use case, the resulting dataset, risk map, will have associated a metadata object that will be created by the process described above. The final step in the workflow is when the user decides to publish the metadata record to a catalogue service the metadata manager checks the validity of the resource associated metadata, the validation will depend on the metadata standard that has been chosen to publish, thus the standards that the Catalogue Service supports. If the metadata conform to minimum requirements according to the selected output metadata standard format, then the metadata manger uses stylesheets to generate an XML string compatible with the catalogue service and carries out the pertinent interaction with the server to establish the connection and publish the metadata.

[24] http://www.inspire-geoportal.eu

This prototype is only capable of working with shapefile[25] vector layer file format. The implemented and supported metadata standard format is the core of the ISO19115:2003 standard. The implementation of the transformation templates for this format has been made based on the standard specification document published by ISO (ISO/FDIS19115). However, its architecture has been designed to support all the desired functionality. So, somehow, this is a proof of the concept of the complete functionality that will be captured in the future Metadata Extension.
Using this integrated solution, the user can close the life cycle of metadata (Baca, 2008) within the same application. So we could create, modify and publish metadata using the metadata manager, and later discover and recover metadata using the integrated catalog client in gvSIG that allow us to recover the linked data from the Catalog Service.

6. Conclusions and Future Work

Metadata descriptions are critical to enhance the discovery, access and use of GI data and therefore are a key element in achieving good data integration and smooth functioning of Spatial Data Infrastructures, as a basic infrastructure to discover, share and use heterogeneous GI data. This points out the need to facilitate metadata production to easily create, with minimal user intervention, metadata descriptions in standard formats.
The presented methodology includes mechanisms capable of automatic generation and publication of metadata in Open Catalogues as means of improving geospatial information sharing in distributed environments like SDI.
As a proof of concept the implemented prototype allows the extraction of explicit metadata from data resources to be catalogued for data discovery in a Spatial Data Infrastructure.
This implementation of the concept of semiautomatic extraction and management of metadata facilitates the creation and edition of images and geospatial data to be published in a Spatial Data Infrastructure. The integrated nature of this solution within the user workflow hopefully will lead to a proliferation of metadata creation, thus improving the functionality and value of SDIs. Furthermore, this development completely supports the philosophy of total integration of data and metadata that we are trying to promote in order to all data generated are easily found and accessible.
In the early future we will complete the development of the metadata manager for documenting well-known imagery and cartographic data sources. The work includes document more types of resources and file formats, add new standard formats and expand the possibilities of publication, but the major effort will be done to continue implementing and improving each of the methods that compose the proposed methodology to automatically generate metadata included in the *Metadata Generation Engine*. Furthermore, this metadata generation engine is a generic approach, so it may be extended to include new data types and multimedia content.
As we had said, the automatic metadata generation methodology includes more intelligent methods to extract metadata by using inferential reasoning techniques from other metadata and data associated. Intuitive extraction of intrinsic (context-based) metadata of the data source in Google-like techniques, including deductive methods to create well formed free text.

[25] http://en.wikipedia.org/wiki/Shapefile

Another interesting future development is the *GeoCrawler*, a massive metadata generation application, that using crawler techniques and the proposed automatic metadata generation methodology, will allow us to automatically describe the resources available in old data collections currently without documenting, or simply in the user local machine. Subsequently, these data may be published or indexed with respect to the information contained in their metadata to be easily found and accessible by other users. We also consider very interesting the possibility of use this kind of crawler applications in user's local machines, allowing them to share their multimedia resources automatically, for example, in the current social networks.

On the other hand, we will continue to investigate and develop new techniques that allow us the complete integration of data and their metadata. This will greatly facilitate the management, reuse and sharing of resources. Additionally, we will explore other lines of investigation about georreferenced resources publication, for example, the use of indexing techniques that allows us to find the data using simple metadata sets, rather than creating complex formats such as those stored in the current catalogs.

7. Acknowledgements

This work was partially funded by the project "CENIT España Virtual", funded by the CDTI in the program "Ingenio 2010" through *Centro Nacional de Información Geográfica* (CNIG).

8. References

Antoniadis, P., Le Grand, B., 2007. Incentives for resource sharing in self-organized communities: From economics to social psychology. In *Digital Information Management*, 2007. ICDIM '07

Baca, M, 2008. "Introduction to Metadata: Pathways to Digital Information (version 3.0)". In Getty Research Institute.

Bodoff, D., Hung, P.C.K, Ben-Menchem, M., 2005. Web metadata standards: observations and prescriptions. In IEEE Software, January-February 2005, pp. 78-85.

Boutell M., Luo J. , 2005. Beyond pixels: Exploiting camera metadata for photo classification. In *Pattern Recognition*, v. 38, n. 6, pp. 935-946.

Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., 2000. Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation 6 October 2000. http://www.w3.org/TR/2000/REC-xml-20001006. (last accessed in July 2009)

Bulterman D. Is it Time for a Moratorium on Metadata? IEEE Multimedia, October-December (2004) 10-17.

Cantan-Casbas, O., López-Pellicer, F. J., Nogueras-Iso, J., Zarazaga-Soria, F. J. 2008. Issues hampering the widespread adoption of catalogues based on the OGC Catalogue Services Specication. In *Computers & Geoscience* 2008.

Cerda, D. (2005). El mundo según Google: Google Earth y la creación del dispositivo GeoSemántico global. http://geosemantica.gearth.googlepages.com (last accessed in July 2009)

Craglia, M., Kanellopoulos, I., Smits, P. Metadata: where we are now, and where we should be going. Proceedings of 10th AGILE International Conference on Geographic Information Science 2007. Aalborg University, Denmark

Day, M., Tzong-Han Tsai, R., Sung, C., Hsieh, C., Lee, C., Wu, S., Wu, K., Ong, C., Hsu, W., 2007. Reference metadata extraction using a hierarchical knowledge representation framework. In Decision Support Systems, v. 43, pp. 152–167.

Díaz, L., Martín, C., Gould, M., Granell, C., Manso, M.A. Semi-automatic Metadata Extraction from Imagery and Cartographic data, International Geoscience and Remote Sensing Symposium (IGARSS 2007). Barcelona, Julio 2007. IEEE CS Press, pp. 3051-3052.

Eckerson, Wayne W., 1995. "Three Tier Client/Server Architecture: Achieving Scalability, Performance, and Efficiency in Client Server Applications." Open Information Systems 10, 1 (January 1995): 3(20)

FGDC, 2000. Content Standard for Digital Geospatial Metadata Workbook, version 2.0. Federal Geographic Data Committee (FGDC), Metadata Ad Hoc Working Group.

Goodchild, M. (2008). Assertion and authority: the science of user-generated geographic content. http://www.geog.ucsb.edu/%7Egood/papers/454.pdf (last accessed in July 2009)

Goodchild, M. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal* 69 (4): 10. 0343-2521.

Greenberg, J., Spurgin, K., Crystal, A., 2005. Final Report for the AMeGA (Automatic Metadata Generation Applications) Project. UNC, School of Information and Library Science University of North Carolina.

Hand D., Mannila H., Smyth P, 2001. Principles of Data Mining, Cambridge. The MIT Press.

Hill, L. (2006). Georeferencing. In The MIT Press. ISBN 0-262-08354-6.

Kawtrakul, A., Yingsaeree, C.A., 2005. Unified Framework for Automatic Metadata Extraction from Electronic Document. In *Proceedings of IADLC2005 (The International Advanced Digital Library Conference)*, pp. 71-77, Nagoya, Japan.

Klien, E., Lutz, M., 2005. The Role of Spatial Relations in Automating the Semantic Annotation of Geodata. In *Proceedings of the Conference of Spatial Information Theory (COSIT'05)*, Lecture Notes in Computer Science, v. 3693, pp. 133-148, Ellicottville, NY, USA.

Manola F, Miller E, (eds) (2004). RDF Primer. W3C, W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-primer-20040210. (last accessed in July 2009)

Manso, M.A., Nogueras-Iso, J., Bernabé, M.A., Zarazaga-Soria, F, 2004. Automatic metadata extraction from geographic information. In *Proceedings of the 7th AGILE conference on* Geographic Information Science, pp. 379-385, Heraklion, Greece.

Masser, I., Rajabifard, A., Williamson, I. Spatially enabling governments through SDI implementation. International Journal of Geographical Information Science. Vol. 22, No. 1, (2008) 5–20

Nebert, D., 2004. Developing Spatial Data Infrastructures: The SDI Cookbook v.2.0. In *Global Spatial Data Infrastructure (GSDI).* http://www.gsdi.org/gsdicookbookindex.asp (last accessed in July 2009)

Nebert, D., Whiteside, A., 2004. OpenGIS – catalogue services specification (version 2.0). OpenGIS Project Document 04-021r2, Open GIS Consortium Inc.

Nogueras-Iso, J., Zarazaga-Soria, F.J., Béjar, R., Álvarez, P.J., Muro-Medrano, P.R. OGC Catalog Services: a Key element for the development of Spatial Data Infrastructures, Computers and Geosciences, vol. 31/2, (2005) 199-209.

Rajabifard, A., Feeney, M-E.F., Williamson, I. P. Future directions for SDI development. International Journal of Applied Earth Observation and Geoinformation 4 (2002) 11–22

Rüdiger Schollmeier, 2002. A Definition of Peer-to-Peer Networking for the Classification of Peer-to-Peer Architectures and Applications. In Proceedings of the First International *Conference on Peer-to-Peer Computing, IEEE*.

Sanchidrian Cano, N., Calle González, J.V., 2005. Editor de Metadatos NEM v 1.0 para ArcGis. In *Jornadas Técnicas de la IDEE de España (JIDEE05)*, Madrid.

Stewart, C and Kowaltzke, A. 1997, Media: New Ways and Meanings (second edition), *JACARANDa*, Milton, Sydney. pp. 102.

Suh, B., Bederson, B.B., 2007. Semi-Automatic Photo Annotation Strategies Using Event Based Clustering and Clothing Based Person Recognition. In *Interacting With Computers*, v. 19, n. 4, pp. 524-544. Elsevier.

Taussi, M., 2007. Automatic production of metadata out of geographic datasets (master's thesis). University of Technology, Department of Surveying. Helsinki, Espoo. http://www.tkk.fi/Units/Cartography/theses/master/2007/Diplomityo_Taussi_M.pdf (last accessed in July 2009)

Toebes John, 2007. Enabling a Richer Video Experience With Metadata. A position paper for the W3C Video on the Web Workshop. 12-13 December 2007, San Jose, California and Brussels, Belgium. Chief Architect, Cisco Media Solutions Group. Available in http://www.w3.org/2007/08/video/positions/Cisco_MSG.html (last accessed in July 2009)

Zarazaga-Soria, F.J. Lacasta, J., Nogueras-Isso, J., Torres, M.P., Muro-Medrano, P.R., 2003. A Java Tool for Creating ISO/FGDC Geographic Metadata. In *Geodaten- und Geodienste-Infrastrukturen - von der Forschung zur praktischen Anwendung. Beiträge zu den Münsteraner GI-Tagen*. IfGI prints. 2003, vol. 18, pp. 17-30.

IntechOpen

Geoscience and Remote Sensing

Edited by Pei-Gee Peter Ho

ISBN 978-953-307-003-2
Hard cover, 598 pages
Publisher InTech
Published online 01, October, 2009
Published in print edition October, 2009

Remote Sensing is collecting and interpreting information on targets without being in physical contact with the objects. Aircraft, satellites ...etc are the major platforms for remote sensing observations. Unlike electrical, magnetic and gravity surveys that measure force fields, remote sensing technology is commonly referred to methods that employ electromagnetic energy as radio waves, light and heat as the means of detecting and measuring target characteristics. Geoscience is a study of nature world from the core of the earth, to the depths of oceans and to the outer space. This branch of study can help mitigate volcanic eruptions, floods, landslides ... etc terrible human life disaster and help develop ground water, mineral ores, fossil fuels and construction materials. Also, it studies physical, chemical reactions to understand the distribution of the nature resources. Therefore, the geoscience encompass earth, atmospheric, oceanography, pedology, petrology, mineralogy, hydrology and geology. This book covers latest and futuristic developments in remote sensing novel theory and applications by numerous scholars, researchers and experts. It is organized into 26 excellent chapters which include optical and infrared modeling, microwave scattering propagation, forests and vegetation, soils, ocean temperature, geographic information , object classification, data mining, image processing, passive optical sensor, multispectral and hyperspectral sensing, lidar, radiometer instruments, calibration, active microwave and SAR processing. Last but not the least, this book presented chapters that highlight frontier works in remote sensing information processing. I am very pleased to have leaders in the field to prepare and contribute their most current research and development work. Although no attempt is made to cover every topic in remote sensing and geoscience, these entire 26 remote sensing technology chapters shall give readers a good insight. All topics listed are equal important and significant.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Arturo Beltran, Laura Diaz, Carlos Granell, Joaquin Huerta and Carlos Abargues (2009). Description and Publication of Geospatial Information, Geoscience and Remote Sensing, Pei-Gee Peter Ho (Ed.), ISBN: 978-953-307-003-2, InTech, Available from: http://www.intechopen.com/books/geoscience-and-remote-sensing/description-and-publication-of-geospatial-information

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China

www.intechopen.com

51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen

IntechOpen

© 2009 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen